CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD
COLLEGE

Admin

- **Note-taker**: Femi
- **Midterm 1 handed out today**
  - Do not open until you are ready to take it!
  - 2 hour time limit
  - Due Thursday at the beginning of class (Oct 7)

Why do we have a exam?

- Process of synthesizing the material on your own is essential

- Preparing the “study sheet” is designed to facilitate that process

- Review in class today and in lab (working through midterm practice questions and notecards)

Outline for September 30

- Review
  - Linear regression
  - Gradient descent
  - Matrix/vector form of Lab 3
  - Classification
  - Single feature models / decision trees
  - Evaluation metrics

PLEASE LEAVE COMPUTERS ON

Matrix/Vector form of SGD

while not converged:

shuffle the data (stochastic)

4, 7, 2, 1, 10 — shuffle the indices instead

for $i = 1, 2, 3 \cdots n$:

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}^{t+1} \leftarrow \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}^{t} - \vartheta \overbrace{\Big(\underbrace{h_{\vec{w}}(\vec{x}_i)}_{\text{prediction}} - y_i \Big)}^{\text{Scalar}} \begin{bmatrix} x_{i0} \\ x_{i1} \\ x_{i2} \\ x_{ij} \\ \vdots \\ x_{ip} \end{bmatrix}$$

test for convergence

x_{ij} is the jth feature of the ith example

row ↑ ↑ col

model

$h_{\vec{w}}(\vec{x}) = \vec{w} \cdot \vec{x}$

fake 1

$= w_0 x_0 + w_1 x_1 \cdots w_p x_p$

$= \sum_{j=0}^{p} w_j x_j$

$$\text{converged} = \underbrace{\left| \text{new cost} - \text{old cost} \right|}_{\text{expensive operation}} < \varepsilon$$

shuffle the data

$i = 1 \ldots n$ (training data)

$\Rightarrow$ need to shuffle X & y together

$p = 1$

$t=1 \quad (x_1, y_1)$ & (x_2, y_2)

$t=2 \quad (x_2, y_2)$ & (x_1, y_1)

cost(X, y, w)

$$J(\vec{w}) = \frac{1}{2} \sum_{i=1}^{n} \underbrace{(\hat{y}_i - y_i)^2}_{\text{residuals}} \Big\} \text{model error}$$

$$= \frac{1}{2} (\vec{\hat{y}} - \vec{y}) \cdot (\vec{\hat{y}} - \vec{y})$$

$$\boxed{J(\vec{w}) = \frac{1}{2} (X\vec{w} - \vec{y}) \cdot (X\vec{w} - \vec{y})}$$

$\boxed{\text{page 3 of handout}}$

$\boxed{\text{predict}(X, w)}$

$\Rightarrow$ return $\hat{y}$ (vector)

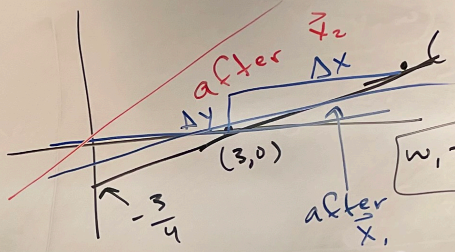
$$\underbrace{n \Big\{ \underbrace{\begin{bmatrix} \\ x_{i0} \; x_{i1} \; \cdots \; \cdots \; x_{ip} \\ \\ \end{bmatrix}}_{p}}_{X} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} = \begin{bmatrix} \vdots \\ \vec{X}_i \cdot \vec{w} \\ \vdots \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{bmatrix} \Big\} \vec{\hat{y}}$$

y_{pred}

(7)

$$X = \begin{array}{c} \\ \\ \end{array}\begin{bmatrix} 1 & 3 \\ 1 & 7 \end{bmatrix}\begin{array}{l} \leftarrow \vec{x}_1^{\,T} \\ \leftarrow \vec{x}_2^{\,T} \end{array}, \quad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

ideal model $= \boxed{h(\vec{x}) = -\frac{3}{4} + \frac{1}{4}x}$



$\boxed{w_1 = \frac{1}{4}}$ (4: $7-3$) $= \frac{\text{rise}}{\text{run}}$

extra: use analytic solution

$w = (X^T X)^{-1} X^T y$

$w = \begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \end{bmatrix}$

$\boxed{i=2}$ $\vec{x}_2 = \begin{bmatrix} 1 \\ 7 \end{bmatrix}, \quad \vec{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1\Big(\underbrace{0}_{\vec{w}\cdot\vec{x}_2} - \underbrace{1}_{y_2}\Big)\begin{bmatrix} 1 \\ 7 \end{bmatrix}$$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix}$$

$\boxed{i=1}$ $\vec{x}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad \vec{w} = \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix}$

$(0.1\cdot 1 + 0.7\cdot 3) = 2.2$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \leftarrow \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} - 0.1\left(\begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix}\cdot\begin{bmatrix} 1 \\ 3 \end{bmatrix} - 0\right)\begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$\boxed{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} -0.12 \\ 0.04 \end{bmatrix}}$$

PLEASE LEAVE COMPUTERS ON

time varying α (SGD)

$t = 1$

while not converged:

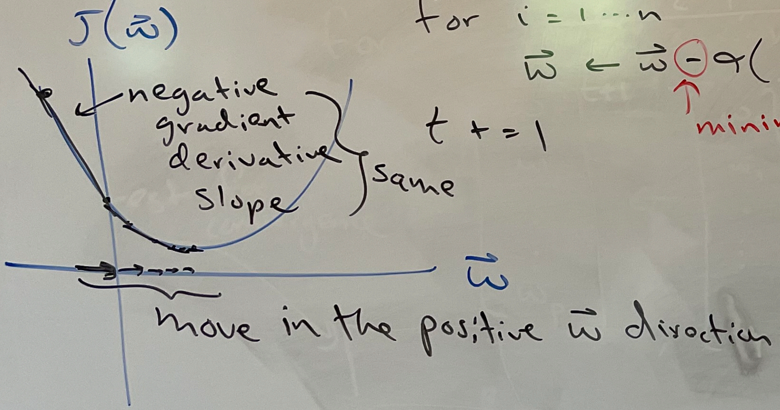
$\alpha = \frac{1}{t}$ # α decreases over time

for $i = 1 \dots n$

$\vec{w} \leftarrow \vec{w} - \alpha(\hat{y}_i - y_i)\vec{x}_i$

minimizing J

$t += 1$

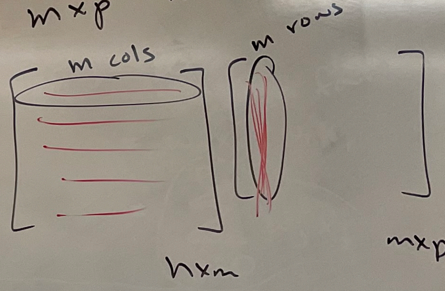


Runtime

$A \Rightarrow n \times m$ matrix

$B \Rightarrow m \times p$ matrix

$\underset{n \times m}{A}\ \underset{m \times p}{B}$? =



$\cancel{\underset{m \times p}{B}\ \underset{n \times m}{A}}$

how long for dot product?

$\boxed{O(m)}$

multiple linear regression (p features)

$$h_{\vec{w}}(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 \cdots w_p x_p$$

polynomial regression (p = 1)

$$h(x) = w_0 + w_1 x + w_2 x^2 + \cdots w_d x^d$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & & & & \\ 1 & & & & \\ 1 & & & & \\ 1 & & & & \end{bmatrix}$$

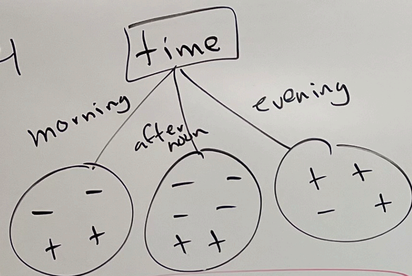
Lab A

Computer Science

DO NOT ERASE
purgatory
sad machines (3x)
yao
all H110, fix as needed
purchase & con
turing or church
3 - check prices 2TB
DO NOT ERASE

Nasanbayar
CS260 TA Hours:
6:30-8:30pm

#4



1 model, many thresholds

0.5	+	−	+
0.75	−	−	+
0.1	+	+	+

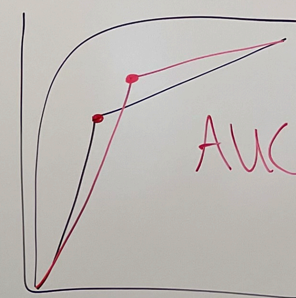
→ error $= \frac{7}{14}$

training data → evaluation metrics

error : $\frac{2+2+1}{14} = \boxed{\frac{5}{14}}$, accuracy = 1 - error $= \frac{9}{14}$

pred

true \ pred	−	+
−	4	3
+	2	5



AUC = area under curve

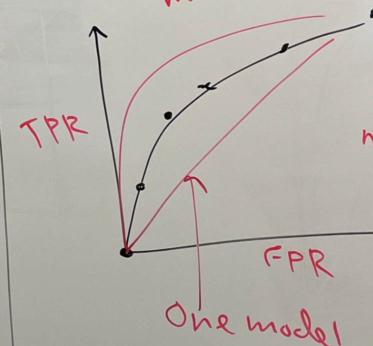
Computer Science

1. predict $\underbrace{\{\text{History, Physics, Art}\}}_{\text{class labels}}$ → multi-class classification
2. predict height → regression
3.
   - internal nodes — feature names (i.e. time)
   - branches — feature values i.e. morning, afternoon, evening
   - leaves — class labels → poisonous, edible

5. (a)

	−	+
−	42	8 "false alarm"
+	8 "miss"	12

- $FPR = \frac{8}{50} = 0.16$ → one x-value
- $TPR = \frac{12}{20} = 0.6$ → one y-value

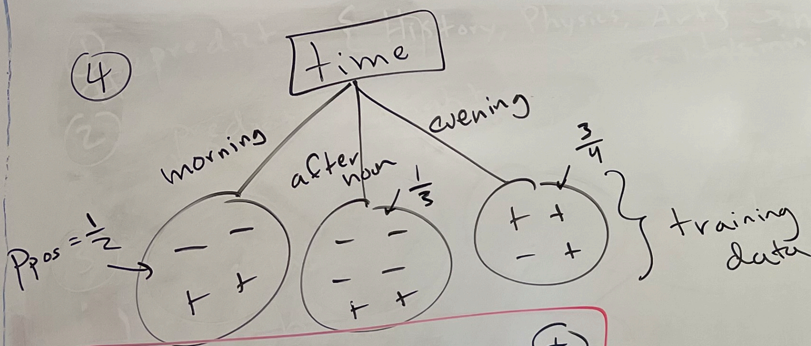


for each feature → make a new model (train)
  for every threshold
  - re-classify test data
  - new conf. mat.
  - 1 pt on ROC

Lab B

Computer Science

4



0.5	+	−	+
0.75	−	−	+

$\underbrace{0.75}_{\text{thresholds}}$ t

=> probabilities

$\boxed{P_{pos} \geq t \Rightarrow +}$

evaluation metrics → usually on test data
can do on training data

error : $\frac{2+2+1}{14} = \boxed{\frac{5}{14}}$

accuracy
$= 1 - \text{error}$
$= \boxed{\frac{9}{14}}$

confusion matrix

	pred −	pred +	
true −	4	3	→ 7
true +	2	5	→ 7

Computer Science

output

(1) predict {History, Physics, Art}

classification ↑ Lab 4

(2) predict height

↑ regression ↳ Lab 2, 3

(3)

- internal nodes & root
- branches
- leaves

class labels output

feature names

feature values

time

values

(5)

	pred −	pred +
−	42	8
+	8	12

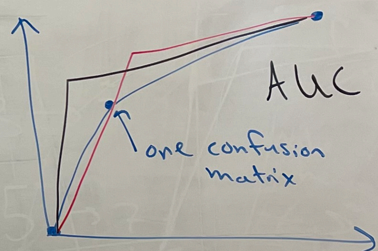
false alarm

miss

higher ⟷ lower thresh

$$FPR = \frac{8}{50} = 0.16$$

$$TPR = \frac{12}{20} = 0.6$$



AUC : area under the curve

Sam A
CS3340 TA
7:30-9:30
Red Shirt

← Sam

← Keeton

Keeton M
CS 107 TA
till 8:20 or so

Xavier D
CS107 TA
11:30-12:30
Blue Shirt

Freddie
Econ 300 TA
8-9:30

	−	+
−	28 + 56	0
+	11 + 4	0

[84, 15]

age ≤ 55.5

F: [28, 11] (− +) → −

T: [56, 4] (− +) → −

$$\text{error} = \frac{11+4}{84+15} = \boxed{\frac{15}{99}}$$

[84, 15]

old peak ≤ 3.55

[0, 2] → +

[84, 13] → −

[0 0] → −

$$\text{error} = \frac{13}{99}$$

	−	+
	84	0
	13	2

Sam A
CSB340 TA
7:30-9:30
Red Shirt

 ← Sam

 ← Keeton

Keeton M
CS 107 TA
till 8:20 or so

Xavier D
CS 107 TA
11:30-12:30
Blue Shirt

Freddie
Econ 300 TA
8-9:30

⑦ analytic

$$X = \begin{bmatrix} 1 & 3 \\ 1 & 7 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{matrix} 1 & 1 \\ 3 & 7 \end{matrix}$$

$$\hat{w} = (X^T X)^{-1} (X^T y)$$

$$= \left(\begin{bmatrix} 1 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 1 & 7 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 1 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$$

$$\left(\begin{bmatrix} 2 & 10 \\ 10 & 58 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 7 \end{bmatrix}$$

$$\rightarrow \hat{w} = \begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$